Classification of Literary Style that Takes Order into Consideration

Xavier Puig, Martí Font, Josep Ginebra[1]

The statistical analysis of the heterogeneity of the style of a text often leads to the analysis of contingency tables of ordered rows. When multiple authorship is suspected, one can explore that heterogeneity through either a change-point analysis of these rows, consistent with sudden changes of author, or a cluster analysis of them, consistent with authors contributing exchangeably, without taking order into consideration. Here an analysis is proposed that strikes a compromise between change-point and cluster analysis by incorporating the fact that parts close together are more likely to belong to the same author than parts far apart. The approach is illustrated by revisiting the authorship attribution of *Tirant lo Blanc*.

KEY WORDS: Authorship attribution, Bayesian analysis, Stylometric analysis, Model based cluster, Correlated data, Word length, Function words.

1 Introduction

The statistical analysis of literary style has often been used to settle authorship attribution problems both in the academic as well as in the legal context. Early work used word length and sentence length to characterize literary style. Other characteristics widely used for this purpose have been the proportion of nouns, articles or adjectives, the frequency of use of function words, which are independent of the context, and the diversity of the vocabulary used by the author. As a consequence, data in this context is almost always categorical.

[1]Xavier Puig and Martí Font are Lecturers and Josep Ginebra is Professor at the Departament of Statistics, E.T.S.E.I.B., Universitat Politècnica de Catalunya, Avgda. Diagonal 647, 6^{a} Planta, 08028Barcelona, Spain (E-mails: xavier.puig@upc.edu, josep.ginebra@upc.es).

In the particular case where one suspects that there might be more than one author, one typically carries out an heterogeneity analysis of the style of the text or corpus of texts after splitting it down into smaller pieces. Under most of the stylometric characteristics listed above, that leads to the analysis of a contingency table that will often have ordered rows, with each row corresponding to a different piece of the text or corpus, and each column corresponding to the counts of a given category, like of a function word or words or sentences of a given length.

One approach to that problem is through single change-point analysis, assuming that the ordered rows share style and hence the same distribution all the way up to a given point of the row sequence, where the author changes and hence the style and that distribution changes and stays the same for the remaining sequence of rows in the table. The goal in that type of analysis is estimating both the change-point, as well as the before and after the change-point distributions that help characterize the differences in style between authors. This naturally generalizes to multiple change-point analysis, and it is useful in settings where one can assume that the change of author has been sudden.

An alternative approach is through cluster analysis, also recognized as unsupervised classification, which consists on partitioning the rows of the table into groups that are more homogeneous than the whole and could be sharing the same style, without imposing any order restriction when forming the groups. That approach can be implemented based on finite mixture models and it is useful when authors can be assumed to be intervening exchangeably.

Between change-point analysis that force all consecutive observations except the ones at change-points to belong to the same group, and cluster analysis, that assign observations to groups without taking order into consideration, there is a whole range of analysis that incentive but do not force consecutive observations to belong to the same group. That fits

well the authorship attribution settings where one is willing to assume that consecutive parts are more likely to belong to the same author than parts that are far apart.

Here one such analysis is proposed based on an extension of the finite mixture models that incorporate the fact that the role of authors could be changing along the text. By letting neighboring observations be related, the model will also capture the correlation that one expects to find as a consequence of the way the writing process works.

Most of the alternative classification methods that are used in the literature of authorship attribution and of the analysis of the heterogeneity of literary style assume data to be continuous, when in practice most of the time data is categorical. We avoid that continuity assumption. Furthermore, the usual classification methods employed by the authorship attribution literature use ad hoc heuristic partitioning algorithms that tend to be easy to apply and work well, but do not allow one to assess cluster uncertainties and do not provide rigorous inference based methods to allocate individual observations to clusters, (see, e.g., Kaufman and Rousseeuw, 1990, Gnanadesikan, 1997, or Gordon, 1999).

Instead, in this manuscript Bayesian model based clustering approaches are adopted, under which observations are assumed to come from one of two sub-populations, each with a distinctive distribution. These approaches provide a complete probabilistic framework assuming a finite mixture model under which observations (texts) belonging to the same cluster (author) have the same distribution, and then estimating the mixed distributions and assigning observations to these distributions. Each one of the two distributions involved in the mixture characterize each one of the two styles. Model based approaches simultaneously group objects and estimate the distribution of each group, and that avoids the biases appearing whenever these two stages are tackled separately.

Model based Bayesian methods also have the advantage over the usual heuristic classification methods of providing a measure of the uncertainty in the allocation of individual

observations into clusters, and of casting the decision of the number of clusters (authors) as a statistical testing problem. Good introductions to Bayesian and non Bayesian model based classification methods can be found in Bock (1996), McLachlan and Peel (2000) and Fraley and Raftery (2002).

To illustrate our novel approach, the authorship attribution problem of *Tirant lo Blanc* will be revisited by analyzing the *word lengths* and the use of *function words* in its chapters, and the results will be compared with the ones of the change-point and cluster analysis of this data carried out in Giron, Ginebra and Riba (2005).

The paper is organized as follows. Section 2 presents the authorship attribution problem that will be used to illustrate the method and motivate its need. In Section 3 the model proposed is presented and compared with the multinomial change-point and cluster models. In Section 4 the results of the analysis for *Tirant lo Blanc* is presented, and in Section 5 possible extensions are discussed.

2 Description of the authorship problem

Tirant lo Blanc is a chivalry book written in catalan, hailed to be "the best book of its kind in the world" by Cervantes in El Quixote, and considered by many to be the first modern novel in Europe, (see, e.g., Vargas Llosa, 1991, 93). The main body of the book was written between 1460 and 1464, but it was not printed until 1490, and there has been a long lasting debate around its authorship, originating from conflicting information in its first edition.

Where in the dedicatory letter at the beginning of the book it is stated that "*So that no one else can be blamed if any faults are found in this work, I, Joanot Martorell, take sole responsibility for it, as I have carried out the task singlehandedly*," in the colophon at the end of the book it is stated that "*Because of his death, Sir Joanot Martorell could only*

finish writing three parts of it. The fourth part, which is the end of the book, was written by the illustrious knight Sir Martí Joan de Galba. If faults are found in that part, let them be attributed to his ignorance." Over the years, experts have split between the ones defending the existence of a single author for all its 487 chapters, in line with the dedicatory letter, and the ones backing a change of author somewhere between chapters 350 and 400, in line with the colophon. For a detailed overview of this debate, see Riquer (1990).

It is well accepted by all medievalists that the main (and maybe single) author, Joanot Martorell, died in 1465, and did not start work on the book before 1460, and that if there were any additions, they would be close to the end of the book and made by the second author much later, when the book was printed in 1490. Neither Martorell nor the candidate to be the book finisher left any other texts comparable with this one.

An analysis of the diversity of the vocabulary carried out in Riba and Ginebra (2006) finds that it becomes significantly less diverse after chapter 383. Giron et al (2005) carried out a multinomial change-point analysis and a multinomial two-cluster analysis based on word lengths and on the frequency of words that do not depend on context, called function words; under both characteristics a stylistic boundary is detected between chapters 371 and 382, apparently with a few chapters misclassified by that boundary. Section 3.1 describes and motivates these two types of analysis. As in these previous studies, here the edition of *Tirant lo Blanc* by Riquer is used; after excluding from consideration the titles of chapters, the quotations in latin and the chapters with less than 200 words, that leads to the analysis of a total of 398242 words split down into 425 chapters.

The literature on the statistical analysis of style characterized through word length and through the use of function words is far too large to be covered in detail here. Early uses of word length can be found for example in Mendenhall (1887), Mosteller and Wallace (1984), Brinegaar (1963), Bruno (1974), Williams (1975), Morton (1978), Smith (1983) and Hilton

Word length counts												
Chapter	1	2	3	4	5	6	7	8	9	10+	N_i	$\overline{wl}_i$
1	21	59	44	19	33	20	16	17	9	17	285	4.47
2	53	113	80	49	52	33	28	36	16	16	476	4.14
...	...	...	...	...	...	...	...	...	...	...	...	...
487	48	49	62	53	41	36	21	9	16	13	348	4.20
Function word counts												
Chapter	e	de	la	que	no	l	com	molt	és	jo	si	dix
1	12	15	9	8	1	7	2	1	6	0	3	0
2	26	28	19	9	3	2	3	8	3	1	3	1
...	...	...	...	...	...	...	...	...	...	...	...	...
487	29	13	8	10	2	10	3	9	0	0	0	0

Table 1: Part of the 425×10 table of word length counts in chapters of more than 200 words of *Tirant lo Blanc*, and of the 425×12 table of counts of twelve function words in them. N_i is the total number of words and $\overline{wl}_i$ is the average word length. Authors will provide the full data set to anyone requesting it.

and Holmes (1993). Early uses of function words can be found in some of these references as well as in Burrows (1987, 92), Holmes (1985, 92), Binongo (1994) or Oakes (1998). Function words are proven to be more sensitive than word length when trying to tell authors apart.

In the example of *Tirant lo Blanc* the analysis of word length leads to the analysis of the 425×10 table of word length counts partially presented in Table 1, and the analysis of the twelve function words used in Giron et al. (2005) leads to the analysis of the 425×12 table of function words partially presented in that table. These twelve function words were chosen in that paper by first doing a change-point and a cluster analysis of the chapters of the book based on the 25 most frequent words, and then selecting the subset of these words that best discriminated between the estimated two groups of chapters.

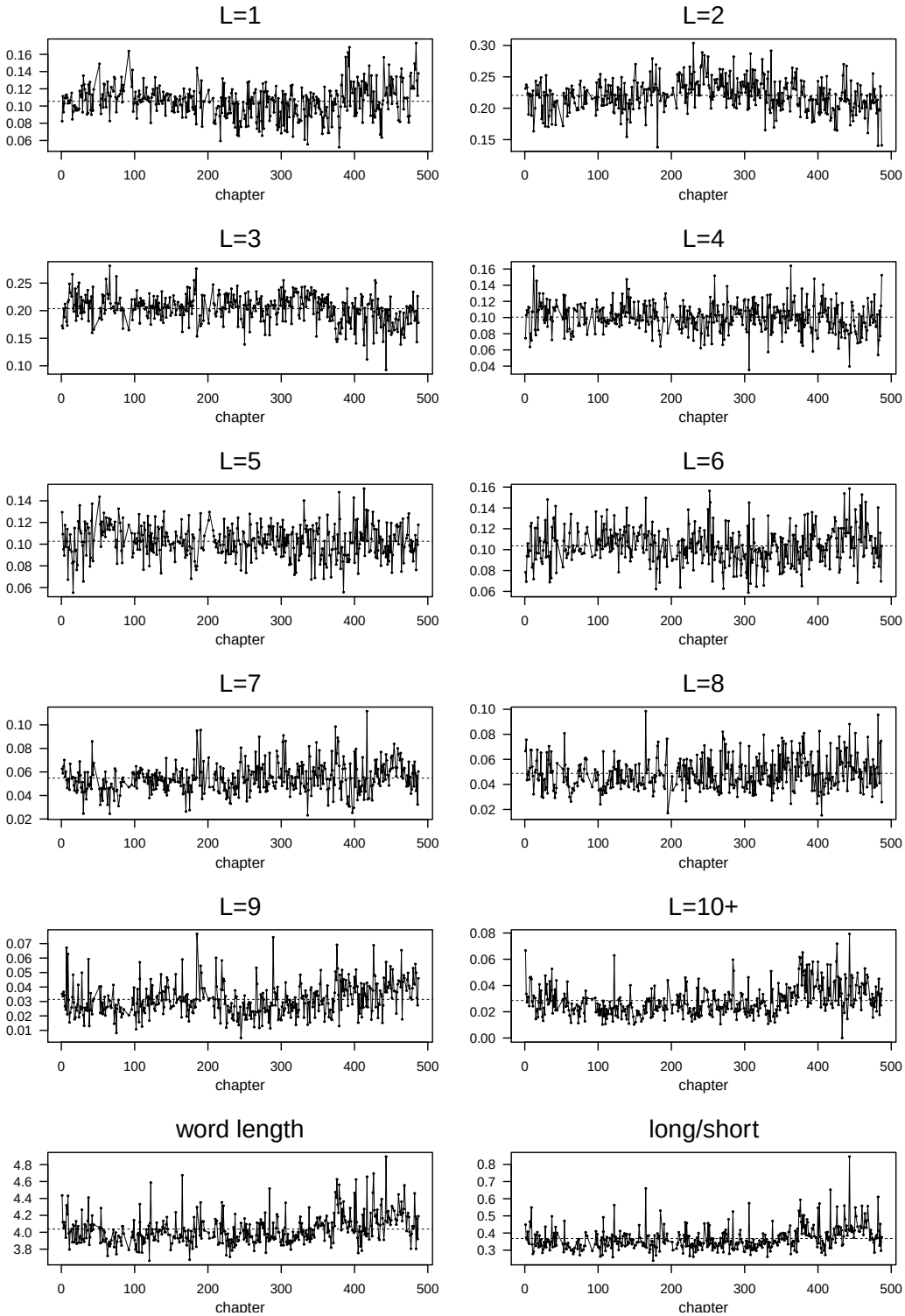

Figure 1: Sequence of proportion of words of each length in each chapter of *Tirant lo Blanc*, with $L = l$ meaning words of l characters, sequence of average word length, and sequence of the ratio between the number of long words and of short words in them.

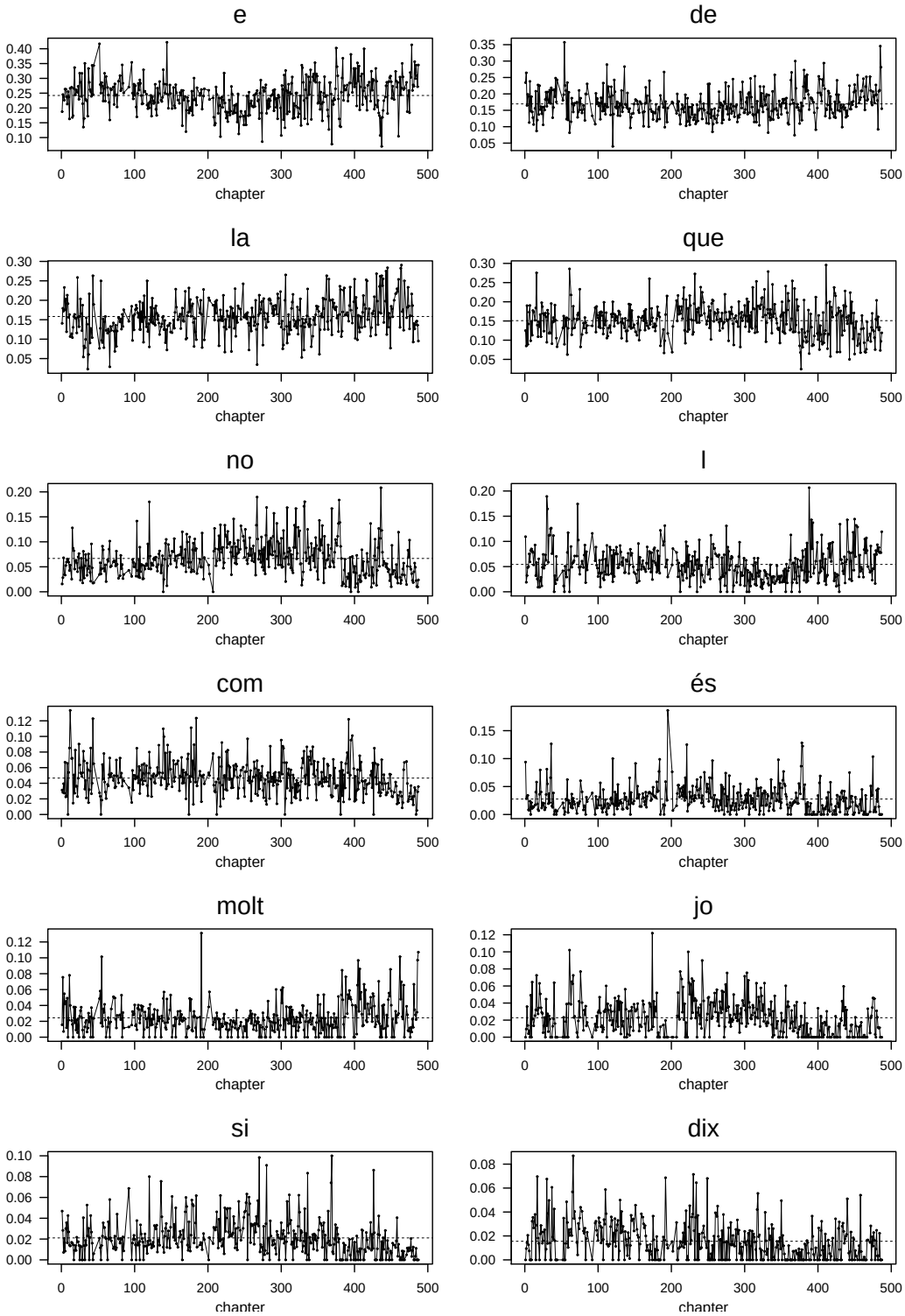

Figure 2: Frequency of appearance in the chapters of *Tirant lo Blanc* of the twelve function words used in the analysis.

If the book had been written by a single author, one might expect the proportion of words of each length and the frequency of use of each function words to be similar in all chapters. As a consequence, one would expect that once taken into account the fact that chapters have different lengths, all the rows in each one of the two sub-tables of Table 1 would have similar distributions. If instead, the distribution of these rows either changed suddenly or kept switching back and forth between two different distributions, it could indicate the existence of a second author that either took over at some point and completed the book, or contributed chapters all over the book.

Figure 1 presents the sequence of the proportions of words of each length in each chapter, the sequence of the average word length per chapter and the sequence of the ratio between the number of long words, (with six or more letters), and of short words, (with less than six letters). Note that, for example, the average word length and the proportion of single lettered words and of ten or more lettered words seems to be larger at the end of the book. Figure 2 presents the sequence of frequencies of the twelve function words selected. Note that there is also a clear shift in the level of use of words like *e, que, no, l, molt, jo* or *dix* towards the end of the book. What is found in both figures might be consistent both with the existence of two authors and a single change-point, as well as with the existence of a second author filling in material mostly at the end of the book.

In some instances, one might explain changes in style through differences in chronology or topic, specially when one is dealing with works that were written during a long span of time. In our case though it is known that the main author (single author according to some) of the book worked on the book during a short span of time, shortly before his death, and therefore in our example differences in style should not be attributed to breaks in writing. That the estimated changes in style do not coincide with shifts in topic needs to be checked after the chapters are classified according to style.

The three models considered next assess whether the observations in these sequences can be adequately classified into two different populations, each corresponding to a different style. The first model assumes that the change happens once suddenly, the second model assumes that the two styles alternate exchangeably all over the text, and the third model strikes a compromise somewhere in between.

3 Description of the models

For each chapter in the book (or part in the corpus of texts), i with $i = 1, \ldots, n$, one has a vector valued categorical observation, $y_i = (y_{i1}, \ldots, y_{ik})$, where k denotes the number of categories of the stylistic characteristic. In our example, y_i will be the ten dimensional vector of word length counts in the i-th chapter, presented as the i-th row in the first sub-table of Table 1, and the twelve dimensional vector of frequency counts of the function words selected in that chapter, presented as the i-th row in the second sub-table. The set of all the rows in each sub-table will be denoted by $y = (y_1, \ldots, y_n)$.

Under all the three models considered next, the i-th row of the table, y_i, will always be assumed to be multinomially distributed, $\text{Mult}(N_i, \theta_i)$, where $N_i = \sum_j^k y_{ij}$ denotes the i-th row total and hence the total number of words considered in that row, and where $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})$ is such that $\sum_{j=1}^k \theta_{ij} = 1$, with θ_{ij} being the probability of the j-th category for the i-th row. In our example k will be ten for the first table of word lengths and twelve for the second table of function words. Thus, the rows of these two tables will be considered to form sequences of conditionally independent observations with probability density function (pdf):

$$\text{Mult}(y_i|N_i, \theta_i) = \frac{N_i!}{\Pi_{j=1}^k y_{ij}!} \Pi_{j=1}^k \theta_{ij}^{y_{ij}}. \tag{3.1}$$

The vector of probabilities, $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})$, can be seen as a fingerprint of the style of the author in his texts, because one expects that on average he will use different categories

of words with the same relative frequencies. That will lead to the texts by the same author sharing the same set of average probabilities, θ_i. Under that assumption, θ_i characterizes the style of the author while N_i naturally takes into account the text size and therefore the weight to be allocated in the analysis to each row of each table.

If all the chapters belong to the same author and were written at about the same time, it is reasonable to expect that they will share the same style and therefore one would expect the vector of probabilities, θ_i, for all the rows in the two sub-tables considered to stay approximately constant along the whole sequence of 425 chapters. In that case, the rows of these sub-tables could be modeled as a random sample of $\text{Mult}(N_i, \theta)$ distributions.

On the other hand, if one detects a sudden shift in the vector of probabilities, θ_i, through a change-point analysis, that might indicate a sudden change in style and therefore a sudden change of author, of topic, or of writing time. If, instead, one identifies the rows of the tables as belonging to two distinct populations through a cluster analysis, with each population of rows sharing a different vector of probabilities, that might indicate the existence of two different styles and therefore of two different authors intervening more or less exchangeably all along the book. Next, these two settings are modeled probabilistically.

3.1 Multinomial change-point and cluster models

In a multinomial single change-point analysis one assumes that $y = (y_1, \ldots, y_n)$ is a sequence of conditionally independent multinomial random variables such that $\theta_i = \theta_b$ for $i \leq r$ and $\theta_i = \theta_a$ for $i > r$, and thus with a probability density function (pdf):

$$p(y|r, \theta_b, \theta_a) = \prod_{i=1}^{r} \text{Mult}(N_i, \theta_b) \prod_{i=r+1}^{n} \text{Mult}(N_i, \theta_a). \tag{3.2}$$

This model assumes that the first r chapters (rows) before the change-point have been written by the first author with a style characterized by the first set of probabilities θ_b, while the remaining set of $n - r$ chapters (rows) after that change-point have been written

by the second author with a style characterized by the second set of probabilities θ_a. The goal in change-point analysis is to learn about the change-point, r, as well as about the before and after the change-point multinomial probabilities, θ_b, θ_a, characterizing the two styles.

As an alternative, in multinomial two-cluster analysis, the n rows of the table, $y = (y_1, \ldots, y_n)$, are considered to be conditionally independent and identically distributed according to a finite mixture of two multinomial distributions, with pdf:

$$p(y|\omega, \theta_1, \theta_2) = \prod_{i=1}^{n} (\omega * \text{Mult}(N_i, \theta_1) + (1 - \omega) * \text{Mult}(N_i, \theta_2)), \tag{3.3}$$

where $\theta_s = (\theta_{s1}, \ldots, \theta_{sk})$ for $s = 1, 2$ determine the distribution of the rows in the s-th cluster, and hence characterize the style in that cluster, and where ω is a weight determining the proportion of rows belonging to the first cluster and hence the probability that any given row will be allocated to that cluster. This model assumes that the chapters (rows) allocated to the cluster 1 were written by an author with a style characterized by the set of probabilities θ_1, while the remaining chapters (rows) allocated to the cluster 2 were written by a different author with a style characterized by θ_2.

To allocate rows into clusters, which is an essential feature in cluster analysis, one has to introduce a vector of unobserved (latent) categorical variables $\zeta = (\zeta_1, \ldots, \zeta_n)$, where ζ_i takes values in $\{0, 1\}$ and is such that $\zeta_i = 1$ when the i-th row belongs to the first cluster and $\zeta_i = 0$ when it belongs to the second cluster. A variable is considered to be latent whenever one can not observe it but is willing to estimate it, very much like one does for a parameter. Here the ζ_i are assumed to be conditionally independent and identically distributed, with $\pi(\zeta_i = 1|\omega) = \omega$ and $\pi(\zeta_i = 0|\omega) = 1 - \omega$. As a consequence the joint pdf for $y = (y_1, \ldots, y_n)$ and $\zeta = (\zeta_1, \ldots, \zeta_n)$ becomes:

$$p(y, \zeta|\omega, \theta_1, \theta_2) = \prod_{i=1}^{n} (\omega * \text{Mult}(N_i, \theta_1))^{\zeta_i} ((1 - \omega) * \text{Mult}(N_i, \theta_2))^{1-\zeta_i}. \tag{3.4}$$

The allocation of rows into clusters can be inferred through point estimates of ζ.

Fitting these multinomial change-point and cluster models through the classical frequentist inference techniques is complicated, specially when it turns to assessing the uncertainty of the estimates of the multinomial probabilities and to estimating ζ. Instead, we adopt the Bayesian inference approach, that requires eliciting a prior distribution on the parameters of the models that summarize the knowledge one has about them, and then updating these distributions in the light of the data. For an introduction to the Bayesian approach to data analysis, see, e.g., Gelman et al. (2013) or Carlin and Louis (2008).

As a prior distribution, one typically assumes by default that the vectors of multinomial probabilities in the change-point analysis, (θ_b, θ_a), and in cluster analysis, (θ_1, θ_2), are independent and Dirichlet$(a_{s1}, \ldots, a_{sk})$ distributed, with either $s = a, b$ or $s = 1, 2$, and hence with pdf:

$$\pi(\theta_s) = \pi(\theta_{s1}, \ldots, \theta_{sk}) = \frac{\Gamma(\sum_{j=1}^{k} a_{sj})}{\prod_{j=1}^{k} \Gamma(a_{sj})} \theta_{s1}^{a_{s1}-1} \ldots \theta_{sk}^{a_{sk}-1}, \tag{3.5}$$

where $\Gamma(\cdot)$ stands for the Gamma function. Depending on the values chosen for $(a_{s1}, \ldots, a_{sk})$, the prior can go from being very subjective to reflecting vague information about the multinomial vector of probabilities, (θ_b, θ_a) and (θ_1, θ_2). In particular, note that the prior expected value for $\theta_s = (\theta_{s1}, \ldots, \theta_{sk})$ will be $(a_{s1}, \ldots, a_{sk})/(\sum_{j=1}^{k} a_{sj})$, and one can chose the a_{sj} to reflect the fact that one knows that some categories have larger probabilities than others. One can also rely on the fact that the larger $\sum_{j=1}^{k} a_{sj}$, the smaller the prior variances of the probabilities θ_{sj}, and hence the more informative the prior will be about θ_s. In the implementation that follows all the $(a_{s1}, \ldots, a_{sk})$ are set to be equal to $(1, \ldots, 1)$, which corresponds to assuming a uniform distribution on the simplex and hence that $E[\theta_{sj}] = 1/k$ for all j, and that all the possible values for $\theta_s = (\theta_{s1}, \ldots, \theta_{sk})$ are equally likely, but more informative distributions have also been tried. In particular note that in the case of function words the categories are ordered from words appearing more frequently to words appearing less frequently, and hence it is also be natural to chose $(a_{s1}, \ldots, a_{sk})$ such that $a_{s1} \geq a_{s2} \geq \ldots \geq a_{sk}$, which lead to $E[\theta_{sj}]$ being decreasing with j.

As a prior distribution for the change-point, r, in the change-point model, one typically chooses a uniform distribution on $\{1, \ldots, n\}$, which assumes that the change in style could happen anywhere in the book equally likely. Nevertheless, if one suspects that the change-point is more likely to happen in certain chapters than in certain others, one should incorporate that information in a more informative prior.

In the cluster analysis model, as a prior for the cluster weight, ω, which is the probability that any chapter belongs to Cluster 1 and therefore takes values between 0 and 1, one typically assumes it to be Beta(b, c) distributed and independent of (θ_1, θ_2), which is a very flexible family of distributions supported on $[0, 1]$ with pdf:

$$\pi(\omega) = \frac{\Gamma(b+c)}{\Gamma(b)\Gamma(c)}\omega^{b-1}(1-\omega)^{c-1}, \tag{3.6}$$

where, again, $\Gamma(\cdot)$ stands for the Gamma function. In the implementation (b, c) is set to be equal to $(1, 1)$, which is the same as assuming that ω takes a uniform distribution on $[0, 1]$, and hence that all possible values for ω are equally likely. For more details on the Dirichlet and Beta distributions, see Johnson, Kemp and Kotz (2005) and Johnson, Kotz and Balakrishnan (1997).

Note that beta and Dirichlet probability models are the default Bayesian choices as prior distributions when one needs to model proportions and vectors of probabilities, respectively. We also tried more informative priors, incorporating the fact that the categories in the second sub-table are ordered from more frequent to less frequent function words. More informative priors for r and ω were also tried, but sample sizes are large enough so that data is so much more informative than any of the prior distributions used and hence the posterior distributions were insensitive to the choice of prior distribution. Hence these distributional choices have very limited impact on the results of the analysis presented in Section 4. For more technical details on these multinomial change-point and cluster models, see Giron et al. (2005).

3.2 Multinomial cluster model with dependence

When carrying out a cluster analysis based on (3.4) one assumes that all rows and corresponding allocation variables, (y_i, ζ_i) for $i = 1, \ldots, n$, are conditionally independent and identically distributed. As a consequence, one is implicitly assuming that the two styles mix exchangeably along the text, without taking into consideration the order in which rows appear, which most often runs against what one anticipates to be happening.

One extension of the finite mixture model in (3.3) that corrects for that, first considered by Fernandez and Green (2002) in the context of Poisson mixtures for spatially indexed data, lets the weights in the mixture vary from row to row, $\omega = (\omega_1, \ldots, \omega_n)$, which leads to:

$$p(y|\omega, \theta_1, \theta_2) = \prod_{i=1}^{n} (\omega_i * \text{Mult}(N_i, \theta_1) + (1 - \omega_i) * \text{Mult}(N_i, \theta_2)), \tag{3.7}$$

where $\omega_i = (\omega_{i1}, \omega_{i2} = 1 - \omega_{i1})$ is such that $0 < \omega_{i1} < 1$, and hence to the rows of the table, $y = (y_1, \ldots, y_n)$, becoming conditionally independent but not identically distributed. As a consequence of that modification, the probability that the i-th row is allocated to the first cluster, ω_i, will be changing from row to row and the set of latent allocation variables, $\zeta = (\zeta_1, \ldots, \zeta_n)$, indicating whether each row belongs to cluster 1 or 2, will be conditionally independent but not identically distributed, with $\pi(\zeta_i = 1|\omega) = \omega_i$ and $\pi(\zeta_i = 0|\omega) = 1 - \omega_i$. The joint pdf of $y = (y_1, \ldots, y_n)$ and $\zeta = (\zeta_1, \ldots, \zeta_n)$ becomes:

$$p(y, \zeta|\omega, \theta_1, \theta_2) = \prod_{i=1}^{n} (\omega_i * \text{Mult}(N_i, \theta_1))^{\zeta_i} ((1 - \omega_i) * \text{Mult}(N_i, \theta_2))^{1-\zeta_i}, \tag{3.8}$$

and the allocation of the i-th row into either one of the two clusters will be done again based on point estimates of ζ_i. The posterior distribution of ω_i is closely related to the one of ζ_i, and it also helps determine the role of the two authors along the text.

A second feature of the basic cluster model in (3.3) that runs against what one anticipates in most authorship attribution settings is that it does not consider rows (chapters) that are close to be more likely to belong to the same cluster (author) than rows (chapters)

that are far apart. Here, certain degree of sequential dependence in chapter authorship is incorporated through a prior structured distribution of the weights, ω_i, making it more likely that rows in nearby locations have more similar allocation probabilities than rows that are located far apart. More specifically, here one will let ω_i be such that its log odds are:

$$\log \frac{\omega_i}{1-\omega_i} = \alpha_i + \beta_i, \tag{3.9}$$

where the α_i's and the β_i's for $i = 1, \ldots, n$ are terms playing a different role each, and are treated as random effects and hence linked by a hierarchical structure that lets their relative contributions be determined by data.

The term α_i is assumed to be conditionally independent and Normal$(\mu_\alpha, \sigma_\alpha^2)$ distributed, and hence with a contribution to the log odds of ω_i that is comparable for all i, thus capturing the global unstructured heterogeneity in ω_i induced by a likely large set of unobserved covariates. The term β_i is assumed to be conditionally independent and Normally distributed, with their mean and variance being equal to $(\beta_{i-1} + \beta_{i+1})/2$ and $\sigma_\beta^2/2$ for $i = 2, \ldots, n-1$, and with mean and variance being equal to β_2 and σ_β^2 for $i = 1$, and being equal to β_{n-1} and σ_β^2 for $i = n$. By relating the mean of β_i, corresponding to the i-th row (chapter) to the values taken by β_{i-1} and β_{i+1} corresponding to the $(i-1)$-th and the $(i+1)$-th rows (chapters), that term captures the local dependence effect that one expects to find when the degree of intervention of the authors shifts smoothly in the book.

The distribution for ω_i chosen here mimics the priors used by the disease mapping literature to obtain spatially smoothed estimates of Poisson means ever since Besag et al (1991) and Mollie (1996). The novelty is that here the prior is used on time and not space indexed data and that it is used to model dependence through the mixing weights of a cluster model and not through the mean parameter of a single cluster distribution. One can think of other ways of inducing sequentially dependent allocations of rows into clusters, but as long as they are flexible enough and use enough information about neighboring observations, they

should all lead to similar results.

Fitting this model to the data through classical frequentist inference tools would be extremely difficult, and that is why here again the Bayesian inference approach is adopted. That requires one to chose a prior distribution on the parameters of the model to start with, and then compute the posterior distribution by incorporating the information in the data.

If the prior distributions chosen have little information compared with the information in the data, as it will be the case in our implementation, the choice of prior distribution barely has any influence on the posterior distribution, and hence on the inferences reached. Hence, in that case one can think of the choice of a prior distribution as a default technical step where one only needs to be careful to match the parameter set with the support of the priors chosen.

Here, as a prior distribution for μ_α, the expected value of the α_i, one assumes that it is $Normal(m, s)$ distributed, centered at the value expected for the average of the log odds for ω_i, which in our example will be $m = 0$, and with a large variance, that in our example will be set to be $s = 100$. By choosing a normal distribution with a large variance, one is assuming that one knows very little about the mean of the α_i and hence the inferences about these parameters will be very weakly influenced by the choice of that prior.

The inverse of σ_α^2 and of σ_β^2 are non-negative real valued, and by default they are typically assumed to be Gamma(c, d) distributed, and hence to have a pdf:

$$\pi(\sigma) = \frac{d^c}{\Gamma(c)} \sigma^{c-1} e^{-d\sigma}. \tag{3.10}$$

In the implementation that follows one chooses $c = 1$ and $d = .01$, which correspond to assuming that the distributions for σ_α^2 and for σ_β^2 have large variances, which is the standard choice when one wants to use prior distributions that assume that very little is known about

$$(y_1, \ldots, y_n)|\theta_1, \theta_2, \zeta \sim \prod_{i=1}^{n} \text{Mult}(N_i, \theta_1)^{\zeta_i} \text{Mult}(N_i, \theta_2)^{1-\zeta_i},$$

$$(\theta_1, \theta_2) \sim \prod_{j=1}^{2} \text{Dirichlet}(a_{j1}, \ldots, a_{jk}),$$

$$(\zeta_1, \ldots, \zeta_n)|(\omega_1, \ldots, \omega_n) \sim \prod_{i=1}^{n} \text{Bernoulli}(\omega_i),$$

$$\omega_i = e^{\alpha_i+\beta_i}/(1 + e^{\alpha_i+\beta_i}), \quad i = 1, \ldots, n$$

$$(\alpha_1, \ldots, \alpha_n)|\mu_\alpha, \sigma_\alpha^2 \sim \prod_{i=1}^{n} \text{Normal}(\mu_\alpha, \sigma_\alpha^2)$$

$$\beta_1|\beta_2, \sigma_\beta^2 \sim \text{Normal}(\beta_2, \sigma_\beta^2)$$

$$\beta_i|\beta_{i-1}, \beta_{i+1}, \sigma_\beta^2 \sim \text{Normal}((\beta_{i-1} + \beta_{i+1})/2, \sigma_\beta^2/2), \qquad i = 2, \ldots, n-1,$$

$$\beta_n|\beta_{n-1}, \sigma_\beta^2 \sim \text{Normal}(\beta_{n-1}, \sigma_\beta^2),$$

$$\mu_\alpha \sim \text{Normal}(m, s)$$

$$\sigma_\alpha^{-2} \sim \text{Gamma}(c_\alpha, d_\alpha)$$

$$\sigma_\beta^{-2} \sim \text{Gamma}(c_\beta, d_\beta)$$

Table 2: Bayesian multinomial two-cluster model with dependence.

σ. Hence, that choice barely influences the conclusions of the analysis.

As a prior distribution for the multinomial probabilities, (θ_1, θ_2), one assumes that they are independent and with each $\theta_s = (\theta_{s1}, \ldots, \theta_{sk})$ with $s = 1, 2$ having again a Dirichlet$(a_{s1}, \ldots, a_{sk})$ distribution with a pdf as in (3.5). In the actual implementation that follows the $(a_{s1}, \ldots, a_{sk})$ are also set to be equal to $(1, \ldots, 1)$, which corresponds to a reference uniform distribution on the simplex and hence to treating all k categories symmetrically and assuming that all possible values for $\theta_s = (\theta_{s1}, \ldots, \theta_{sk})$ are equally likely. For the details on this default choice as a distribution for (θ_1, θ_2), and for alternative choices that are more informative, we refer to the discussion at the end of Subsection 3.1. Even though the model in (3.7) and (3.8) is more general than the one in (3.3) and (3.4), the role played by these parameters is basically the same in both cases.

The whole Bayesian model, including both the statistical model as well as the prior distributions described above, can be found summarized in Table 2.

An extensive sensitivity analysis has been carried out by trying priors that incorporated different information about the parameters of the hyper prior and of the multinomial parameters. Here it is also found that data is so much more informative than the priors used, that the posterior distribution barely changes by changing the prior choices.

The posterior distribution for the parameters of these models are too complex to be computed analytically. Instead of that, to update the model and simulate from it the WinBugs MCMC implementation has been used (see, Lunn et al. 2000). The convergence of the chains has been assessed through the visual inspection of the sample traces and the monitoring of various diagnostic measures. The authors will provide the code and the data of the example to anyone that requests them.

3.3 Selection of the Number of Authors and Testing

Under each one of the three models contemplated above, that is, the change-point model in (3.2), the cluster model in (3.4), and the cluster model with dependence in (3.8), one needs to chose between the single author (style) case and the two authors (styles) case. In all these situations, that issue can be posed as a choice between two models, and hence can be answered through a formal statistical hypothesis test.

In the change-point model, for example, one needs to test whether $r = n$ (single author) or $r \neq n$ (two authors), and in the basic cluster model, one needs to test whether $\omega = 1$ (single author) or $\omega \neq 1$ (two authors). Resorting to a Bayesian analysis has the advantage that one can select the model with the largest posterior probability. The posterior probability that the M_r model is the one generating the data is:

$$P(M_r|y) = \frac{P(M_r)P(y|M_r)}{\sum_{r=0}^{S} P(M_r)P(y|M_r)}, \tag{3.11}$$

where $P(M_r)$ is the prior probability of model r and where $P(y|M_r)$ is the marginal likelihood of M_r. When one is only interested in comparing models M_r and M_s, one resorts

to:

$$\frac{P(M_r|y)}{P(M_s|y)} = \frac{P(M_r)}{P(M_s)} \frac{P(y|M_r)}{P(y|M_s)}. \tag{3.12}$$

In general, one will select the model with the largest posterior probability; when each model is considered equally likely a priori, the larger the marginal likelihood of a model, $P(y|M_S)$, the more attractive that model.

Most often, computing $P(y|M_S)$ exactly is too complicated to be attempted in practice, but one can estimate $P(y|M_S)$ through the MCMC simulations used to update the model, (see, e.g., Gelfand and Dey 1994 or Raftery and Newton, 1995), which is what will be used next to choose between single and multiple author hypotheses.

4 Results of the analysis of Tirant lo Blanc

Here the word length and the function word data in Table 1 is analyzed using the two-cluster model with dependence just presented, and the result of that analysis is compared with the results obtained using the change-point and basic cluster model in Section 3.1.

A single change-point analysis based on the model in (3.2) leads to a posterior distribution of the change-point, r, highly concentrated around Chapter 371 for the word length data, and highly concentrated around Chapter 382 for the function word data. That explains why the top panels of Figures 3 and 4 assign chapters to authors the way they do. Under both the word length as well as under the function words case, one finds that the posterior probability of the single author (no change-point) model is basically zero; As a consequence, Subsection 3.3 indicates that one should reject the single author hypothesis. Under both tables, the sequence of rows clearly have a change in distribution, indicating a change in style, somewhere between Chapters 371 and 382 of the book.

Under both the basic cluster model in (3.4) as well as the cluster model with dependence in

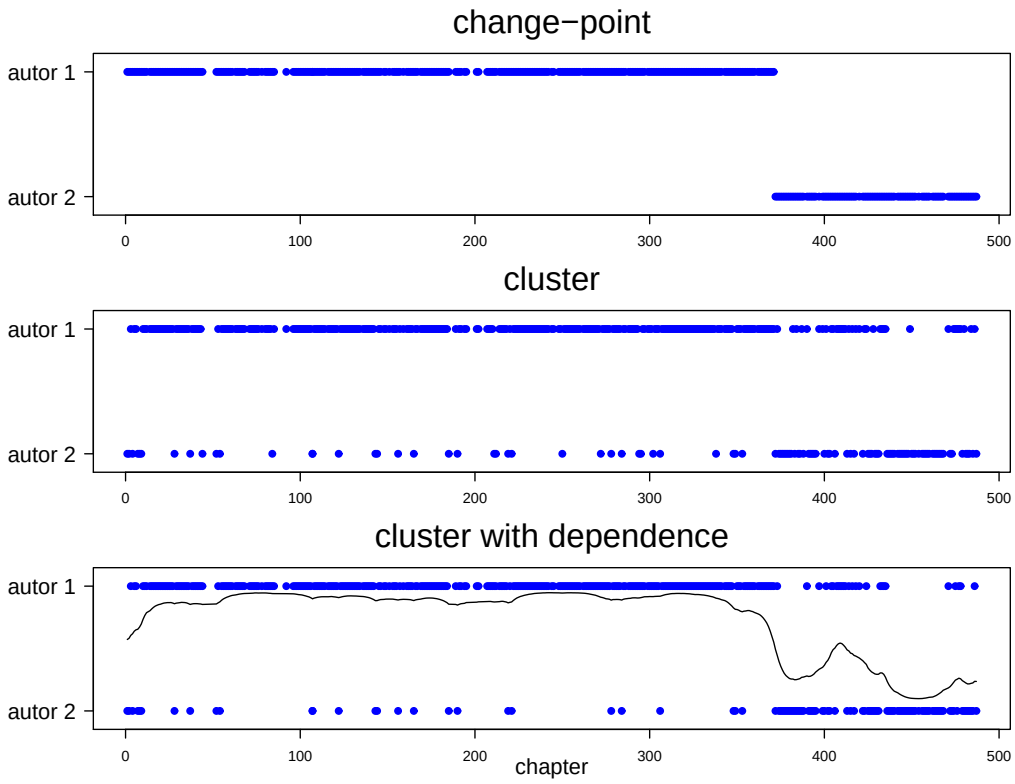

Figure 3: Chapter classification for word length under the single change-point model and under the two-cluster models with and without dependence. The curve on the bottom panel is the posterior expectation of ω_i, which helps describe the role of author 1 in that part of the book.

(3.8), the posterior probability that y_i belongs to the first cluster, $E[\zeta_i|y]$, can be estimated through the MCMC simulated samples. Given that $E[\zeta_i|y]$ can be interpreted to be the probability that the i-th chapter belongs to cluster (author) 1, it is natural to allocate that chapter to cluster (author) 1 whenever $E[\zeta_i|y] > .5$, and to allocate that chapter to cluster (author) 2 otherwise.

The second panel in Figures 3 and 4 presents the classification of chapters into authors according to this rule under the basic cluster model in (3.4). Using word length data, Figure 3 indicates that 319 chapters are attributed to the first author, which represents 75.06% of the 425 chapters considered, and only 75 chapters are classified differently than

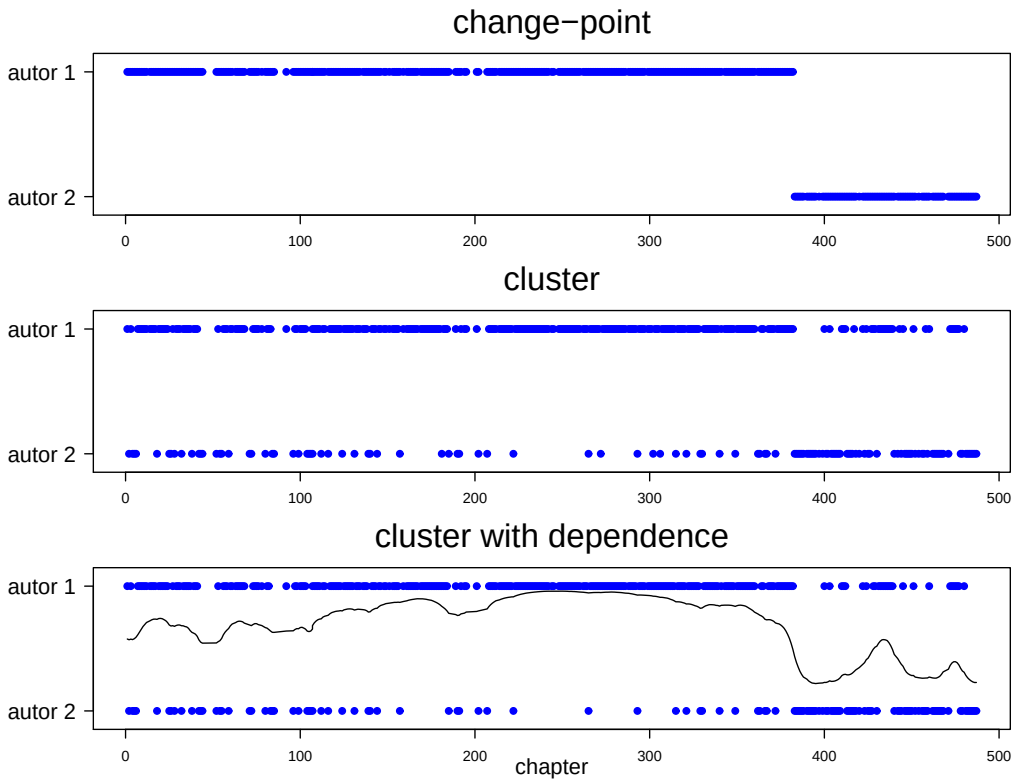

Figure 4: Chapter classification for the function word data under the single change-point model and under the two-cluster models with and without dependence. The curve on the bottom panel is the posterior expectation of ω_i, which helps describe the role of author 1 in that part of the book.

through the change-point model, of which 38 are attributed to the second author but are located before chapter 371, while 37 are attributed to the first author but are located after that chapter. For the function word data, in Figure 4 one finds 304 chapters attributed to the first author, which represents 71.53% of the total; in this case, 59 chapters are attributed to the second author but located before chapter 382, while 32 chapters are attributed to the first author but located after it. When one tests the single author hypothesis against the double author hypothesis, using the idea described in Subsection 3.3, one finds that under both tables the probability of the two-authors hypothesis is almost one, and therefore one again clearly rejects the single author hypothesis.

The third panel in Figures 3 and 4 presents the chapter classification based on the $E[\zeta_i|y]$ under the cluster model with dependence in (3.8). The classification under this more sophisticated model is similar to the one obtained through the basic cluster model, and the corrections are in the direction of making the classification more similar to the one obtained through the change-point model. For the word length data here only 23 chapters are classified differently than through the basic cluster model, with only 27 chapters located before chapter 371 and yet attributed to the second author, and only 25 chapters located after that chapter and yet attributed to the first author. Using function word data only 9 chapters are classified differently than through the basic cluster model, with 56 chapters being attributed to the second author but located before chapter 382 and 28 chapters being located after that chapter but attributed to the first author.

According to the model with dependence, the chapters located before the $371-382$ change-points that are consistently allocated to Author 2 instead of Author 1 under both stylometric characteristics are chapters $2, 4, 28, 52, 54, 107, 144, 185, 190$ and 349 while the chapters located after these change-points that are consistently allocated to Author 1 are $410-412, 424, 432-435, 475$ and 477.

The posterior expected value of ω_i, in the third panel of Figures 3 and 4, also helps describe the role of each author along the book. Whether $E[\omega_i|y]$ is larger or smaller than .5 serves as an indication of which author plays the main role in that part of the book. Note the close agreement between $E[\omega_i|y]$ and the classification of chapters into authors according to the change-point model. This tool is unavailable under the basic two-cluster model.

Once the existence of two authors is settled and chapters are allocated into each one of the styles according to each one of the models, the question arises as to how do the components in $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})$ change when one switches from one style to the other according to each one of the models. To address that, Figures 5 and 6 plot a sample of the posterior

distribution of $\log(\theta_{bj}/\theta_{aj})$ under the change-point model in (3.2) and of $\log(\theta_{1j}/\theta_{2j})$ under the cluster models in (3.4) and in (3.8). Note the high degree of agreement between the three models, and specially between the cluster models with and without dependence, that follows from the agreement in the way these models allocate chapters into styles.

Figure 5 indicates that two, three, four and five lettered words are more abundant in the style of the author writing most of the book, while one, six, seven, eight, nine and ten or more lettered words are more abundant in the style of the author writing mostly at the end of the book. Figure 6 indicates that words *que, no, com, és, jo, si* and *dix* are more abundant in the part of the book written by the main author, while *e, de, la, l* and *molt* are more abundant in the parts of the book written by the second author.

5 Final Comments

The statistical analysis identifies a change in style near chapters 371–382, with a few chapters being misclassified by that change-point. That agrees with the boundary detected in chapter 383 through the analysis of the diversity of vocabulary in Riba and Ginebra (2006), and it is in line with the hypothesis supported by experts attributing more credibility to the colophon of the book than to its dedicatory letter.

The change-point model, (3.2), is very strict in that it assumes that all consecutive chapters (except the r-th and the $(r+1)$-th chapters) belong to the same author, and that will not adapt to most practical settings. The cluster model that does not allow for dependence, (3.4), is more flexible in that it does not take order into consideration when allocating chapters to authors, and that will also fail to model many practical instances. Instead, the cluster model with dependence proposed in (3.8) strikes a compromise somewhere in between, allowing for neighboring chapters to be more likely by the same author without imposing the restriction that they have to be so. Hence the model in (3.8) has the advantage

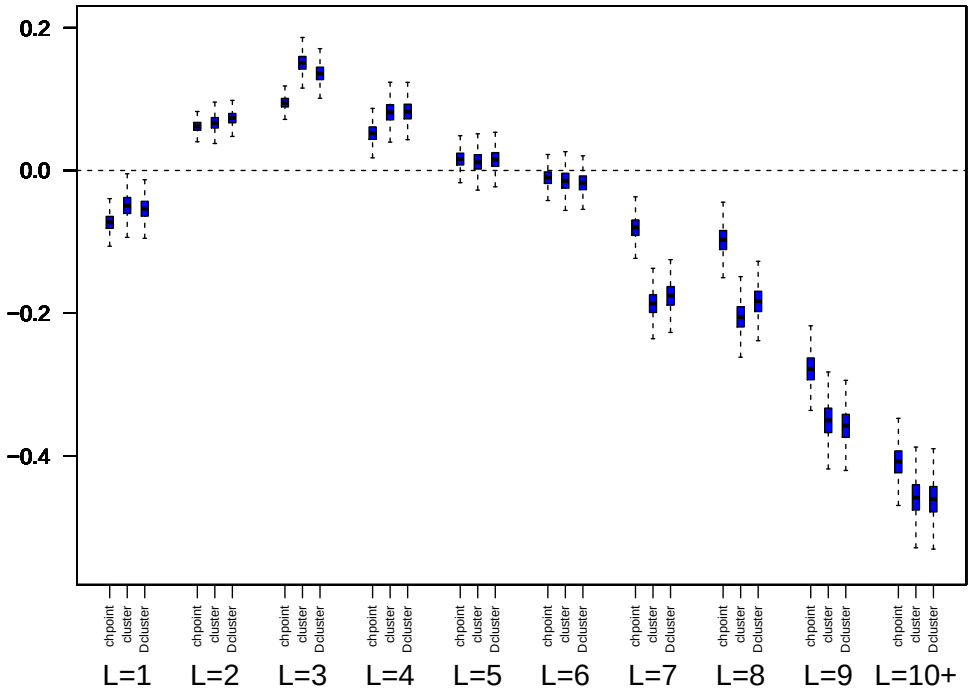

Figure 5: Boxplot of a sample of the posterior distribution of $\log(\theta_{bj}/\theta_{aj})$ under the change-point model, in (3.2), and of $\log(\theta_{1j}/\theta_{2j})$ under the clusters models with and without dependence, in (3.4) and (3.8), for the word length data.

of fitting better the scenarios typically faced in many authorship attribution settings.

As an alternative to the cluster model based on a mixtures of two multinomial models considered here, one could have started with a more flexible framework under which all rows belonging to the same cluster where multinomially distributed with a θ_i that varied from row to row, but with all these θ_i sharing a common distribution. If in particular one assumes that these θ_i are Dirichlet distributed, one would end up basing the analysis on mixtures of two Dirichlet-multinomial models and hence adding two parameters determining the degree of heterogeneity of the multinomial parameters in each cluster. We have tried that approach, but carrying out predictive checks to validate models has lead us to conclude

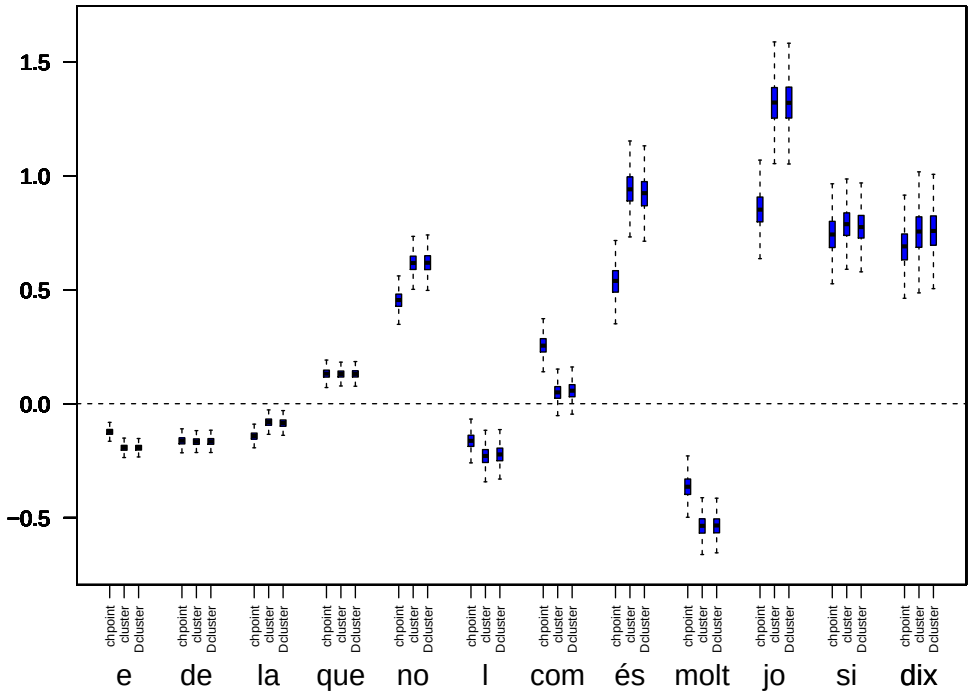

Figure 6: Boxplot of a sample of the posterior distribution of $\log(\theta_{bj}/\theta_{aj})$ under the change-point model, in (3.2), and of $\log(\theta_{1j}/\theta_{2j})$ under the cluster models with and without dependence, in (3.4) and (3.8), for the function word data.

that this type of data does not require these more sophisticated models.

Even though the presentation has focused on the use of word length and function words, and on the two-authors case, it all extends to other stylometric characteristics and to the authorship attribution of texts with more than two authors. A slight modification of the prior for the cluster weights, ω_i, can also accommodate for dependence structures other than the one used here for texts or corpus that are sequentially ordered.

6 Acknowledgments

This work was funded in part by Grant # MTM2010-14887 of the Ministerio de Ciencia e Inovación of Spain. The authors will provide the full data set to anyone requesting it.

7 Bibliography

Besag, J., York, J. and Mollie, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.*, 43, 1-59.

Binongo, J.N.G. (1994). Joaquin's Joaquinesquerie, Joaquinesquerie's Joaquin: a statistical expression of a Filipino writer's style, *Literary and Linguistic Computing*, 9, 267–279.

Bock, H.H. (1996). Probabilistic models in cluster analysis. *Computational Statistics and Data Analysis*, 23, 5-28.

Brinegar, C.S. (1963). Mark Twain and the *Quintus Curtius Snodgrass* letters: a statistical test of authorship, *Journal of the American Statistical Association*, 58, 85–96.

Bruno, A.M. (1974). *Toward a Quantitative Methodology for Stylistic Analysis*, Berkeley, University of California Press.

Burrows, J.F. (1987). Word patterns and story shapes: the statistical analysis of narrative style, *Literary and Linguistic Computing*, 2, 61–70.

Burrows, J.F. (1992). Not unless you ask nicely: the interpretative nexus between analysis and information, *Literary and Linguistic Computing*, 7, 91–109.

Carlin, B.P., and Louis, T.A. (2008). *Bayesian Methods for Data Analysis, 3rd Ed.* Chapman and Hall.

Fernandez, C. and Green, P.J. (2002). Modelling spatially correlated data via mixtures: a Bayesian approach. *J. R. Statist. Soc. B*, 64, 805-826.

Fraley, C. and Raftery, A.E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association,* 97, 611-631.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. (2013). *Bayesian Data Analysis, 3rd ed.* Chapman Hall.

Giron, J., Ginebra, J. and Riba, A. (2005). Bayesian analysis of a multinomial sequence and homogeneity of literary style. *The American Statistician*, 59, 19-30.

Gnanadesikan R. (1997). *Methods of Statistical Data Analysis of Multivariate Observations*, 2nd ed., New York: Wiley.

Gordon A.D. (1999). *Classification*, 2nd ed., London, Chapman and Hall.

Hilton, M.L. and Holmes, D.I. (1993). An assessment of cumulative control charts for authorship attribution, *Literary and Linguistic Computing*, 8, 73–80.

Holmes, D.I. (1985). The analysis of literary style. A review, *Journal of the Royal Statistical Society, Ser A*, 148, 328–341.

Holmes, D.I. (1992). A stylometric analysis of Mormon scripture and related texts, *Journal of the Royal Statistical Society, Ser. A*, 155, 91–120.

Johnson, N.L., Kemp, A.W., and Kotz, S. (2005). *Univariate Discrete Distributions.* New York, Wiley.

Johnson, N.L., Kotz, S., and Balakrishnan, N. (1997). *Discrete Multivariate Distributions.* New York, Wiley.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data*, New York: Wiley.

Martorell, J, and Galba, M.J. (1490). *Tirant lo Blanc* (in catalan), ed. M. Riquer 1983, Barcelona: Edicions 62. (Translated into English by D. Rosenthal in 1986, Baltimore, Johns Hopkins University Press, and by La Fontaine in 1993, Boston, Peter Lang).

McLachlan, G.J., and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS – A Bayesian modelling framework: concepts, structure and extensibility. *Statistics and Computing*, **10**, 325-337.

Mendenhall, T.C (1887). The characteristic curves of composition, *Science*, IX, 237–249.

Mollie, A. (1996). Bayesian mapping of disease. In *Markov Chain Monte Carlo in Practice* (eds W.R. Gilks, S. Richardson and D.J. Spiegelhalter), pp. 359-379. London: Chapman and Hall.

Morton, A.Q. (1978). *Literary Detection*, New York: Scribners.

Mosteller, F. and Wallace, D.L. (1984). *Applied Bayesian and Classical Inference; the Case of The Federalist Papers*, 2nd edn, Berlin: Springer-Verlag.

Oakes, M.P. (1998). *Statistics for Corpus Linguistics*, Edimburg:Edimburgh University Press.

Riba, A. and Ginebra, J. (2006). Diversity of vocabulary and homogeneity of literary style. *Journal of Applied Statistics*, 33, 729-741.

Riquer, M. (1990). *Aproximació al Tirant lo Blanc* (in Catalan), Barcelona:Quaderns Crema.

Smith, M.W.A. (1983). Recent experience and new developments of methods for the determination of authorship, *Association for Literary and Linguistic Computing Bulletin*, 11, 73–82.

Vargas Llosa, A. (1991). *Carta de Batalla por Tirant lo Blanc*, (in Spanish), Barcelona: Seix Barral.

Vargas Llosa, A. (1993). Tirant lo Blanc: Las palabras como hechos, (in Spanish), in *Actes del Symposion Tirant lo Blanc*, 587–604, Barcelona: Quaderns Crema.

Williams, C. B. (1975). Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon, *Biometrika*, 62, 207–212.